DOCUMENT RESUME

ED 219 415 TM 820 426

AUTHOR Cook, Linda L.; And Others
TITLE A Study of the Temporal Stability of IRT Item Parameter Estimates.
PUB DATE Mar 82
NOTE 49p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Aptitude Tests; Equated Scores; *Item Analysis; *Latent Trait Theory; *Test Items; *Test Reliability
IDENTIFIERS *Item Parameters; Scholastic Aptitude Test

ABSTRACT

Data from the Scholastic Aptitude Test-Verbal (SAT-V), SAT Mathematics (SAT-M), and Achievement Tests in Biology, American History, and Social Studies were used for this study. The temporal stability of item parameter estimates obtained for the same set of items calibrated for different examinees at different times was analyzed. It was believed that greater time lapses in test administrations would result in greater differences between item parameter estimates obtained from test administration data. The type of test probably influences the stability of item parameter estimates. Parameter stability is affected by the fit of the data to the model. Aptitude test items were a better fit to the three parameter model. Stability of item parameter estimates was influenced more by differences in group ability than by the length of time between administrations. The item parameter estimates obtained for aptitude test data (SAT-V and SAT-M) had a higher degree of stability than those estimated for achievement tests. Items should be re-calibrated periodically to ascertain if parameter estimates have remained valid for a particular application and examinee population. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219415

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A Study of the Temporal Stability of IRT Item Parameter Estimates[1,2]

Linda L. Cook
Daniel R. Eignor
Nancy S. Petersen

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. L. Cook

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 426

[1] A paper presented at the annual meeting of AERA, New York, 1982.

[2] The authors gratefully acknowledge the advice provided by Frederic M. Lord and Marilyn S. Wingersky in the preparation of sections of this paper.

A Study of the Temporal Stability of IRT Item Parameter Estimates

Linda L. Cook
Daniel R. Eignor
Nancy S. Petersen
Educational Testing Service

One of the attractive features of item response theory (IRT) models is that, theoretically, the parameters characterizing items are invariant across samples of examinees from the same population. If this is also true in application, a number of distinct advantages accrue, advantages that can't be derived from the use of classical test theory methodology (Hambleton, et al, 1978). In the context of this paper, two of these advantages would be the use of invariant item parameter estimates for item banking and equating, particularly for pre-equating. In order for these advantages to accrue, however, it is essential that item parameter estimates obtained at two different points in time, or under two different conditions, be the same apart from sampling error. Proper use of an IRT model with items administered in a particular context, or at a particular point in time, while assuring invariant item parameters at that point, does not guarantee invariance over further uses of these items. As pointed out by Rentz (1978), the issue of invariance, or lack thereof, is fortunately an empirical one that can be investigated. It is also an issue that practitioners in the field of IRT are paying increasingly more attention to.

A number of factors may contribute to a situation in which item parameter estimates obtained for the same set of items, under different conditions, may differ considerably. What follows is a brief outline of these factors, and the relevant research that has been done. The context in which items are calibrated may contribute to a lack of parameter invariance; Whitely and Dawis (1976) have studied context effects using the one-parameter or Rasch model, Yen (1980) using the one- and three-parameter models, and Kingston and Dorans

(1981a) using only the three-parameter model. The homogeneity or heterogeneity of the item set within which items are calibrated has been studied by Bejar (1980) and Kingston and Dorans (1981b) using the three-parameter model, and discussed in detail by Gustafsson (1980) for the Rasch model. The invariance of Rasch parameter estimates across groups of widely differing abilities has been frequently studied (Slinde and Linn, 1978, 1979; Gustafsson, 1979, 1980; Green and Divgi, 1981). Divgi (1981) has provided a review and critique of this literature. Rentz (1978) and Ridenour and Rentz (1980) have looked at the stability of Rasch parameter estimates over time; Kingston and Dorans (1981b) have looked at similar results for the three-parameter model. Finally, Rentz (1982) studied the invariance of parameter estimates for the one- and three-parameter models where there was an intervening instructional program. It should be noted that in all these cases the fact that parameter invariance cannot be demonstrated is because either an inappropriate IRT model was used to characterize the data or one or more of the assumptions underlying IRT have been violated, be it the unidimensionality assumption, the assumption of local independence, or simply the fact that the samples involved in the parameter estimation process are in reality from different populations.

The research presented in this paper extends upon the work of Rentz (1978), Ridenour and Rentz (1980), and Kingston and Dorans (1981b) in that the focus is on the stability of item parameter estimates when the same items are calibrated on two different samples of examinees who have responded to the items at two different points in time, i.e., the temporal stability of the parameter estimates. The data used were collected from regular administrations of the College Board Admissions Testing Program Scholastic Aptitude Test (SAT) and Achievement Tests. The three-parameter logistic model was used to characterize

the relationship between the underlying trait and performance on an item. Theoretically, the item parameter estimates and resulting item response function should not be affected by when the item was administered. Any discrepancy in item parameter estimates obtained for two different samples or at two different points in time should be due to lack of fit of the model due to population shifts, changes in emphases of school curricula over time, or quite simply, due to errors of estimation. As pointed out by Kingston and Dorans (1981b), IRT provides sample invariant parameter estimates for samples (of the same or different ability) from a single population. Population shifts can cause a change in dimensionality and hence, quite different parameter estimates. Divgi (1981b) has pointed out the need to be perceptive of changes in emphases in school curricula, and the effect that these changes may nave on parameter invariance.

There are a number of distinct reasons for our focus in this paper on temporal stability, and more particularly, on the effects of temporal stability, or lack thereof, on IRT equating results. Within the College Board Division of Educational Testing Service (ETS) where the statistical work is done for the SAT and the Achievement Tests, the IRT work, to date, has involved the equating process. The focus has been on (1) a comparison of the results of IRT equatings of the SAT and Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT) to the results obtained from conventional equating methods (Cook, Dunbar, and Eignor, 1981); and (2) the study of scale stability as it is affected by the use of IRT equating methods (Petersen, Cook, and Stocking, 1981). We are soon to embark on a large scale pre-equating study, and as a natural outgrowth of that study, will begin to build a bank of IRT calibrated SAT items; at present such a bank does not exist. A reasonable first question to examine before performing the pre-equating study is what effect the calibration

of the same items using groups of the same and differing abilities at different points in time will have on the stability of the parameter estimates. This question is important because the results of IRT pre-equating, which will be applied to actual SAT administration candidate data, will be generated from pretest data administered to groups of possibly differing abilities at points in time a good deal prior to the actual SAT administration. (Another question of importance, not addressed in this paper, is the effect of pretest context on the parameter estimates and pre-equating results.)

Since the focus, to date, of the work carried out within the College Board Division at ETS has been on the effects of IRT on the equating process, one of the criteria for evaluation of the data being studied here will be the effects of temporal stability, or lack thereof, on equating results. As the item pool mentioned above is developed, and we begin to use IRT for test development purposes, the major focus should then switch to the careful and routine monitoring of the parameters of individual items, rather than the aggregation of items used in the equating process. Divgi (1981b) has made an important point, however, concerning the assumptions of IRT that has relevance for how we have chosen to study the temporal stability of parameter estimates:

> These assumptions are probably satisfied well enough for applications such as equating of intact tests, where IRT is used to predict properties of a large aggregate of items. Validity of the assumptions becomes more important in applications where one deals with individual items, such as tailored testing, item banking and the study of item bias.

Based on Divgi's comments then, it is likely that while one may observe notable differences in individual parameter estimates, or item response functions, these differences may not be apparent in any meaningful

equating comparisons. Because of this potential problem, in this study we will be looking at both summary indices describing the behavior of parameter estimates for individual items and the effects that any differences in these parameter estimates have, when aggregated, on equating results.

Methodology

Data Sets

Data from the College Board Admissions Testing Program Scholastic Aptitude Test (SAT) and Achievement Tests in Biology and American History and Social Studies were used in this study. What follows is a brief description first of the general nature of these examinations and then of the individual forms being studied.

The SAT consists of six 30-minute sections: two verbal sections, two mathematical sections, one Test of Standard Written English (TSWE), and one experimental section, which is either made up of pretest items or a common item equating test which is used to equate the new test form to an existing form. The two verbal sections contain a total of 85 five-choice items composed of 25 antonyms, 20 analogies, 15 sentence completion, and several reading passages each of which is followed by a set of items based on the passage. Scores are reported for the verbal section (SAT-V) based on all 85 items. The two mathematical sections contain a total of 60 items, comprised of 40 five-choice regular mathematics items and 20 four-choice quantitative comparison items. Scores are reported for the mathematics section (SAT-M) based on all 60 items. (TSWE data was not used in this study.) As mentioned previously, the experimental section either contains pretest items, or common item equating sections, 40 items

(10 items of each type) for SAT-V and 25 five-choice regular mathematics items for SAT-M.

The SAT IRT parameter estimates being examined for temporal stability in this study were drawn from a larger set of final forms and equating sections calibrated for the SAT Scale Drift Study (Petersen, Cook, and Stocking, 1981). In choosing the final forms and equating sections to be examined, an attempt was made to choose forms where the time differential between administrations varied from relatively short to long and the samples used in the calibration process were of both comparable and differing abilities. Table 1 (all tables and figures are in the Appendix) presents the final forms and equating sections chosen for study, the numbers of items, administration dates, sizes of the calibration samples and formula score means and standard deviations. Designations starting with a captial letter refer to operational forms of SAT-V or SAT-M; designations consisting of two lower case letters refer to equating sections. In reference to the actual forms and equating sections chosen, Y3 and fw for SAT-V and Y3 and fx for SAT-M were selected because the samples taking the forms at the two administrations were of differing abilities, as judged by the formula-score means. Equating sections fk for SAT-V and fn for SAT-M were chosen because there was little difference in the ability of the samples but interesting time periods between administrations.

The Achievement Tests in Biology and American History and Social Studies both consist of 100 items administered in a 60-minute time period. The American History test focuses on the history of the United States, but other aspects of the social studies also receive attention: in particular, social studies concepts, methods, and generalizations as they are encountered in the

study of history. The Biology test covers a wide variety of specific topics and also includes questions that require the interpretation of experimental data, understanding of scientific methods and laboratory techniques, and knowledge of the history of biology.

The achievement test IRT parameter estimates being examined for temporal stability in this study were drawn from a larger set of forms calibrated for an achievement test scale drift study, which is presently being conducted at College Board Division of ETS. Table 2 presents the same information as Table 1, but for the achievement tests being studied. In choosing the forms to be examined for Biology, one form, VAC1, was chosen because of a large time lapse between administrations (52 months), and the other, TAC2, was chosen because of a significantly shorter time lapse (16 months) between administrations. For both VAC1 and TAC2, the group taking the form at the later administration date is of higher ability, as judged by formula score means. For American History, both forms YAC2 and AAC were chosen to have more or less the same time lapse between administrations but differences, as compared across forms, in the abilities of the groups taking the forms at the two administrations. For YAC2, the abilities of the groups taking the form, as judged by formula score means, are comparable, while for AAC, the abilities are quite disparate.

IRT Model and Method for Developing a Common Metric

Item response theory (IRT) assumes that there is a mathematical function which relates the probability of a correct response on an item to an examinee's ability. (See Lord, 1980, for a detailed discussion). Many different mathematical models of this functional relationship are possible. The model chosen for this

study was the three-parameter logistic model. In this model, where θ represents an examinee's ability, the probability of a correct response to item i, $P_i(\theta)$, is

$$P_i(\theta) = c_1 + \frac{1-c_i}{1+e^{-1.7a_i(\theta-b_i)}}, \qquad (1)$$

where a_i, b_i, and c_i are three parameters describing the item. These parameters have specific interpretations: b_i is the point on the θ metric at the inflection point of $P_i(\theta)$ and is interpreted as the item difficulty; a_i is proportional to the slope of $P_i(\theta)$ at the point of inflection and represents the item discimination; and c_i is the lower asymptote of $P_i(\theta)$ and represents a pseudo-guessing parameter.

The item parameters and examinee abilities for this study were estimated (calibrated) using the program LOGIST (Wood and Lord, 1976; Wood, et al., 1976). The estimates are obtained by a (modified) maximum likelihood procedure with special procedures for the treatment of omitted items (see Lord, 1974). LOGIST requires as input the responses to a set of items from a group of examinees, coded to reflect items answered correctly, incorrectly, omitted, and not reached. In addition, the user may specify certain restrictions on the data and parameters in order to speed convergence of the iterative procedure.

LOGIST produces as output estimates of the a, b, and c for each item, and θ for each examinee. The metric, chosen arbitrarily for the θ (and b) scale, is such that the distribution of estimates of θ has mean zero and standard deviation one. If two separate LOGIST runs are made for the same items, but different groups of examinees, the resulting parameter estimates will be on different scales. There will be, however, a linear relationship that transforms one scale to the other. For all the forms and equating

sections being considered in this study, the parameter estimates were derived from separate LOGIST runs. Because most of the comparisons to be made in investigating temporal stability require the parameter estimates to be on the same scale, a method for developing a common metric had to be used. The method chosen (Lord and Stocking, 1982) is the most recent method to be used at ETS. Briefly, it works as follows. Letting T stand for transformed, a linear transformation of the form

$$b_T = Ab + B \qquad (2)$$

$$a_T = a/A$$

is found which places form two item parameter estimates on the scale of form one. The A and B of this transformation are chosen to minimize the average squared difference between number right true scores on the common set of items for an arbitrary group of examinees who have taken form one. It should be noted that $c_T = c$, so there is no necessity to transform lower asymptote parameters. This method implicitly makes use of information from all the parameters characterizing an item because number right true scores are used in the minimization process.

Methods for Comparing Parameter Estimates

A variety of methods were used in this study for comparing the parameter estimates obtained for the same items calibrated at the separate time points. Since all but two of these methods require that the parameter estimates be on the same scale, and for the two exceptions the same results should obtain from a comparison of transformed or untransformed parameter estimates, all comparisons were performed on the transformed values. The transformation procedure described in the previous section was used to place all parameter estimates on the same scale.

The following methods were used to make comparisons of the parameter estimates obtained for the same items at the separate time points. The first two methods listed could have been applied to either the untransformed or transformed parameter estimates.

1. Two-way plots of the difficulty and discrimination parameter estimates from the two administrations were obtained. If the parameter estimates are indeed invariant, the swarm of points from a two-way plot of the difficulty or discrimination estimates should lie along the same straight line. A visual inspection of such plots can be quite informative.
2. Correlations were calculated between the two sets of parameter estimates for all data sets under study.
3. Means and standard deviations of the parameter estimates (item difficulty, item discrimination and psuedo-guessing) obtained at the separate time points were calculated.
4. The mean of the mean absolute differences ($\overline{MAD}$) between item response functions was calculated for each data set. For each item, two item response functions exist; the item response functions are based on parameter estimates obtained at the separate time points. Using all individuals in the sample taking the earlier of the two administrations, the absolute difference in the item response functions for each person (i.e., value of θ) was obtained and then averaged. The mean of these averages, computed across all items in a test form, can then be used as a summary statistic.
5. Relative efficiency curves were calculated and plotted. The item parameter estimates for the items in each data set from each of the administrations were used to calculate information curves, and then

the ratios of these information curves at various ability levels were used to calculate relative efficiency curves. The earlier administration of the test was always used as the "baseline test" for relative efficiency comparisons.

6. Finally, to test what effect the lack of temporal stability of the parameter estimates has on equating, the following was done. Using the parameter estimates from the two administrations, a true-formula-score equating of a test to itself was performed (see Lord, 1980, Chapter 13). Scores obtained using parameter estimates from the more recent administration of the test were equated to scores obtained using parameter estimates from the earlier administration. If the parameter estimates are truly invariant, the conversion line relating formula scores obtained from the two sets of parameter estimates should have a slope of one and intercept of zero. This line then forms the criterion against which to judge the actual equating, and in turn, to judge what effect the lack of parameter invariance has on the equating process.

The actual true-formula-score equating performed can be described in the following way. The expected value of an examinee's observed-formula score is defined as his or her true-formula score. For the true-formula score, ξ, we have

$$\xi = \sum_{i=1}^{n} \left[\frac{(k_i+1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right] \quad , \qquad (3)$$

where n is the number of items in the test and (k_i+1) is the number of choices for item i. If we have two tests measuring the same ability θ (or two administrations of the same test), then true-formula scores ξ and η from the two test

administrations are related by the equations

$$\xi = \sum_{i=i}^{n} \left[\frac{(k_i+1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right] \qquad (4)$$

$$\eta = \sum_{j=1}^{m} \left[\frac{(k_j+1)}{k_j} P_j(\theta) - \frac{1}{k_j} \right] . \qquad (5)$$

Clearly, for a particular θ corresponding true scores ξ and η have identical meaning. They are said to be equated.

Because true-formula scores below the chance score level are undefined for the three-parameter logistic model, some method must be established to obtain a relationship between scores below the chance level for the two administrations of the same test to be equated. The approach used for this study (Lord, 1980) was to estimate the mean (m) and standard deviation (s) of below chance level scores for the two administrations to be equated via the following formulas:

$$m = \sum_{i=1}^{n} (c_i(k_i+1)/k_i - 1/k_i) \qquad (6)$$

$$s^2 = \sum_{i=1}^{n} (c_i - c_i^2)(k_i+1)^2/k_i^2 \quad , \qquad (7)$$

where n is the number of items in the test, (k_i+1) is the number of choices for item i, and c_i is the pseudo-guessing parameter for item i; and then to use these estimates to define a linear relationship between below chance level scores for the two administrations by setting means and standard deviations obtained from equations 6 and 7 equal.

In practice, true-score equating is carried out by substituting estimated parameters into equations (4) and (5). Paired values of ξ and η are then computed for a series of arbitrary values of θ.

In addition to comparing the true-formula-score equating line to the line with a slope of one and intercept of zero, equating residuals were also calculated and plotted. For any possible score value, the residual was calculated by subtracting the true-formula score for the earlier of the two administrations from the true-formula score for the more recent administration.

Results

The results of the variety of methods for comparing the item parameter estimates from the two administrations, for each data set, are contained in the tables and figures given in the Appendix of this paper. Within each grouping of tables or figures, the sequence of presentation is always the same; SAT Verbal data is presented first, then SAT Mathematical, and finally data from the Achievement Tests under study.

1. Tables 3-5 present the correlations between the parameter estimates, the means and standard deviations of the parameter estimates from the separate calibrations, and the mean of the mean absolute differences ($\overline{MAD}$) between the item response functions.
2. Figures 1-3 present the plots of the item difficulty parameter estimates. Values for the earlier administration of the test are plotted along the abscissa and those for the more recent administration along the ordinate.
3. Figures 4-6 present the plots of the item discrimination parameter estimates. The data is plotted in the manner described for the item difficulty parameter estimates.
4. Figures 7-9 present the relative efficiency curves, where, in each case, the earlier administration of a particular form/equating section served as the baseline test. Due to the ratio nature of relative efficiency calculations (i.e., the ratio of two very small information values can yield a large relative efficiency), data in the tails of these curves should be disregarded.

5. Figures 10-12 present the true-formula-score equating plots. In each plot, the solid straight line is a line with slope of one and intercept of zero and the dotted line is the actual true-formula-score equating line.

6. Figures 13-15 present plots of the equating residuals for all data sets being studied. For each plot, for any (possible) score value, the residual was calculated by subtracting the true-formula score for the earlier administration from the corresponding true-formula score for the more recent administration. The residuals are connected by the dotted line; the solid line forms a baseline against which to compare the dotted line.

Observations may be drawn from the tables and figures at a variety of levels; certain observations hold across all data sets, certain are pertinent to SAT-V, SAT-M, or the achievement tests, and certain are pertinent to particular plots/indices under study.

Examination of the data presented in Table 3 indicates that the correlations among the item difficulty parameter estimates are reasonable for the SAT-V form and equating sections. Correlations among the discrimination parameter estimates are lower and those among the psuedo-guessing parameters, lower still. The degree of correlation is reflected in the scatter plots of the item parameter estimates shown in Figures 1 and 4. Inspection of Figure 1 indicates that difficulty parameter estimates for the SAT-V form/equating sections are fairly stable. The plots show considerable clustering of the points along the straight line. Some scatter, reflective of the lower correlation coefficient given in Table 3, is evident in the plot of SAT-V Y3 data.

The plots of the item discrimination parameter estimates, given in Figure 4, show a greater degree of scatter than the corresponding plots of item difficulty parameter estimates. Again, it can be seen that the data evidencing the greatest degree of scatter is that for SAT-V Y3.

Plots of psuedo-guessing parameter estimates were not obtained. However, it can be seen from inspection of Table 3, that the correlation among these estimates was lowest for SAT-V fw. In general, considering the correlations among the three parameter estimates, those obtained for SAT-V fk indicated the greatest degree of stability and those obtained for SAT-V Y3 the least.

The value of the mean of the mean absolute differences ($\overline{MAD}$) reported in Table 3, indicates that the greatest degree of stability is exhibited by parameter estimates obtained for SAT-V fk and the least amount of stability for those obtained for SAT-V fw. The relatively large value of $\overline{MAD}$ found for the latter equating section is most probably due to the effect on the statistic of the low correlation among the psuedo-guessing parameter estimates.

Plots of relative efficiency curves for the SAT-V form/equating sections are given in Figure 7. The base test, in each instance, represents the earlier administration. The plots can be interpreted in the following manner. If the curve falls below the horizontal line (representing the base test), the test comprised of item parameter estimates obtained at the more recent administration is less efficient then the test with parameter estimates obtained at the earlier administration. The interpretation is reversed for instances where the curved line falls above the horizontal line. It can be seen from examination of the plots in Figure 7 that for SAT-V Y3, the test consisting of item parameter estimates obtained from the more recent administration is slightly less efficient than the test for which items were characterized using data from the earlier administration. The relationship appears to be reversed for the two equating sections fk and fw. With the exception of a slight dip in the curve below the horizontal line for SAT-V fw, the relative efficiency is greater for the more recent administration for both the equating sections.

Plots of the conversion lines resulting from the true-formula-score equatings for the SAT-V form/equating sections are given in Figure 10. The

plots indicate an almost perfect relationship between scores obtained using parameter estimates from the earlier and more recent administrations. More informative are the plots of equating residuals found in Figure 13. The residuals are the differences between the equated true-formula scores for the earlier and more recent administrations of the form/equating sections. Inspection of the graphs indicates that the largest discrepancies are observed between the equated true-formula scores obtained for SAT-V Y3. It should be noted, however, that the residual plots for the equating sections are not directly comparable to that for form Y3 due to differences in number of items. It is possible that a discrepancy of .5 true-formula-score points for a 40 item test might be comparable to a discrepancy of 1.5 points for an 85 item test. The residual plots for the two equating sections can be compared and indicate that the results of the fk equating are slightly better than those obtained for the fw equating.

Summary statistics, correlation coefficients and values of $\overline{MAD}$ for the SAT-M form/equating sections are presented in Table 4. The pattern of the correlation coefficients for the item parameter estimates is similar to that observed for the SAT-V form/equating sections; i.e., the highest correlation coefficients were obtained for item difficulty parameter estimates and the lowest for estimates of the psuedo-guessing parameter. An exception to this pattern is the correlation coefficient obtained for the item discrimination estimates for SAT-M fw. In general, the correlation coefficients between the item parameter estimates obtained for the SAT-M form/equating sections are higher than those obtained for the SAT-V form/equating sections.

Scatter plots of the item parameter estimates are given in Figures 2 and 5. It can be seen, from examination of Figure 2, that the item difficulty estimates appear to be extremely stable, forming tight clusters along the diagonals of the plots. The scatter plots of the item discrimination estimates

(Figure 5) do not exhibit the same degree of stability as the corresponding plots of the difficulty parameter estimates. Upon closer examination of the individual plots, it appears as though the correlation of the item discrimination estimates for SAT-M fn were effected seriously by a single outlier.

The information presented in Table 4 indicates a fairly high degree of stability for all sets of item parameter estimates. The value of the mean of the mean absolute differences is smallest for SAT-M Y3, however, all of the values of this statistic are quite similar.

Examination of the relative efficiency curves presented in Figure 8 indicates that the efficiency of the tests consisting of parameter estimates obtained from the more recent administrations is very similar to that of the tests for which items were calibrated using data from the earlier administrations for both SAT-M Y3 and SAT-M fx. As noted previously, the tails of the curves should be ignored when interpreting the plots. The only plot that is indicative of any degree of instability is the plot depicting the relative efficiency of the two SAT-M fn administrations.

Figure 11 contains plots of the conversion lines resulting from the true-formula-score equatings. As was the case for the SAT-V equatings, the plots indicate an almost perfect relationship between scores obtained using parameter estimates from the earlier and more recent administrations. Plots of the equating residuals, given in Figure 14, indicate very little discrepancy between the equated true-formula scores for SAT-M Y3 and SAT-M fx for all but a few of the lower raw scores. As previously mentioned, the plots for the 25 item equating sections are not strictly comparable to the plot for the 60 item test form. The plot of equating residuals for SAT-M fn indicates a greater degree of discrepancy among equated true-formula scores than do the plots for the SAT-M Y3 and SAT-M fx equatings. It is quite possible that a difference of .5 true-formula-score points is non-trivial.

Summary statistics, correlation coefficients and values of $\overline{MAD}$ for the Biology and American History and Social Studies Achievement Tests are given in Table 5. As indicated from examination of the information in this table, correlations between item parameter estimates are lower than those obtained for either the SAT-V or SAT-M form/equating sections. However, the same general pattern of correlation coefficients is observed; i.e., item difficulty estimates are the most highly correlated and psuedo-guessing parameter estimates the least.

Scatter plots of the item difficulty parameter estimates for the achievement tests are found in Figure 3. The plots indicate a lesser degree of stability than that observed from the plots of the item difficulty estimates for the SAT-V and SAT-M form/equating sections. The plot for American History and Social Studies Form AAC shows a particular amount of scatter. It should be noted that, for all the achievement tests, several item difficulty estimates fell out of the range of the plots. Only one value (b = 3.131, -20.909) obtained for the American History and Social Studies Form AAC seriously affected the correlation coefficient between the parameter estimates.

Figure 6 contains the scatter plots of the item discrimination parameter estimates for the achievement tests. A considerable amount of scatter can be observed in all the plots. The plot with the most extreme outliers appears to be that for American History and Social Studies Form AAC.

The values of $\overline{MAD}$ reported in Table 5 indicate the greatest degree of parameter estimate stability was attained by Biology Form TAC2 and the least degree of stability by American History and Social Studies Form AAC.

Plots of the relative efficiency curves for the achievement test forms are found in Figure 9. For most of the forms, it appears as though the form based on the parameter estimates from the more recent administration is slightly less efficient than the form based on parameter estimates from the earlier administration. The exception is American History and Social Studies Form YAC2. It is somewhat puzzling that the plot for this form indicates the greatest degree of instability for the parameter estimates. This is somewhat contradictory to the information presented in Table 5.

The equating plots presented in Figure 12 indicate, as did the plots for the SAT-V and SAT-M form/equating sections, a close relationship between the scores obtained using parameter estimates from the earlier and more recent administrations. Plots of the equating residuals, found in Figure 15, are more informative. The largest discrepancies between equated true formula-scores were obtained for the American History and Social Studies forms. The discrepancies for all the achievement tests appear to be greater than those obtained for the SAT-V and SAT-M form/equating sections. As mentioned previously, this observation is somewhat confounded by differences in test length.

To summarize, it appears as though some degree of instability is exhibited by all the item parameter estimates. Parameter estimates obtained for the SAT-M form/equating sections exhibit the greatest degree of stability and those obtained for the achievement tests the least. The equating results were surprisingly good for all of the forms/equating sections examined. This suggests that IRT applications that employ aggregates of item parameter estimates may be somewhat robust, at least to the degree of instability of the parameter estimates examined for this study.

Discussion

The purpose of this study was to examine the temporal stability of item parameter estimates obtained for the same set of items calibrated for different samples of examinees at different points in time. It was hypothesized that greater time lapses between earlier and more recent test administrations would result in greater differences between the item parameter estimates obtained using data from the two administrations. An additional hypothesis was that type of test might influence the stability of the item parameter estimates; i.e., if certain types of test data fit the IRT model better, the estimates should remain more stable when calibrated in a variety of circumstances.

Clearly, the item parameter estimates that exhibited the greatest degree of stability were those obtained for the SAT-M form/equating sections. The least stability was demonstrated by the achievement test item parameter estimates; especially those obtained for American History and Social Studies Form AAC. This is not particularly surprising; parameter stability is affected basically by the fit of the data to the model. It is probably true that aptitude test data is less likely to violate the unidimensionality assumption underlying all IRT models then is the type of data obtained for achievement tests, thus resulting in a better fit of the aptitude test items to the three parameter model.

Clear patterns of temporal stability were not evident for any of the forms/equating sections studied. The greatest degree of stability for the SAT-V form/equating sections was exhibited by the parameter estimates for equating section fk and the least amount by Form Y3. It should be recalled that the time lapse between administrations for the SAT-V form/equating sections was greatest for equating section fk. The time lapse between administrations for Form Y3 and equating section fw was similar and about half that for equating

section fk. A possible explanation for the stability of the parameter estimates obtained from the two administrations of SAT-V fk is the similarity in ability levels of the two groups used to calibrate the items.

All the SAT-M form/equating sections exhibited a high degree of stability among item parameter estimates. The one exception appears to be the item discrimination estimates obtained for equating section fn. As mentioned previously, it can be seen from examination of the scatter plot that a single problem item caused the low correlation between the estimates obtained from the two administrations. It does not appear as though time lapse between administrations is related to stability of parameter estimates. The greatest time lapse was observed for the SAT-M Y3 administrations. Item parameter estimates obtained from these administrations resulted in the smallest value of $\overline{MAD}$. The largest value of $\overline{MAD}$ was obtained for equating section fx. The discrepancy between the ability levels of the samples of examinees from the two administrations of this equating section is slightly greater than that observed for SAT-M Y3 or SAT-M fn. Therefore, it appears as though an effect, similar to that observed for SAT-V data, is also observed for these data; i.e., the stability of the item parameter estimates is influenced more by differences in group ability than by length of time between administrations.

The influence of differences in ability level on the stability of parameter estimates suggested by the analyses of the SAT-V and SAT-M forms/equating sections becomes apparent from examination of the data obtained for the achievement tests. The data for these tests indicate a strong relationship between stability of parameter estimates, as assessed by the correlations between the estimates, the scatter plots and the values of $\overline{MAD}$, and discrepancies between ability levels of the samples from the earlier and more recent administrations.

In contrast, length of time between administrations appears to have little affect on the stability of the estimates.

It is possible to draw several conclusions from the information obtained for the various forms and equating sections studied. First, stability of parameter estimates is probably related to type of test. It seems as though parameter estimates obtained for a mathematical aptitude test are more likely to exhibit stability than those obtained for an achievement test in Biology or American History and Social Studies.

Secondly, the stability of the item parameter estimates appears to be more closely related to differences in group ability than to lapses of time between administrations of a test. The important point to note is that for the particular forms/equating sections studied, ability differences appeared to be somewhat unrelated to time differences between administrations. This may not be typical for many testing situations; a situation could easily occur where ability differences would be directly related to length of time between administrations of a test. This could be brought about, for example, by changes in curricular emphases.

The results of the analyses of the equatings were somewhat encouraging. The largest discrepancy in equated-true-formula scores was two points, observed for the American History and Social Studies Form AAC. A discrepancy of two formula score points would result in a discrepancy of approximately 10 reported score points for this test. Although not trivial, the discrepancies are well within the range of the measurement error for the test.

It would appear as though the degree of instability observed in the parameter estimates for the particular forms studied did not impact greatly on the equating results. One important question, not addressed in this study, is the affect of changes in the parameter estimates on the stability of the test scales over time; i.e., is it possible for the small discrepancies observed in

the equatings to accumulate over time, resulting in an upward or downward drift in the test scales.

Also not addressed in the study are the implications of the degree of instability in item parameter estimates for uses besides test equating. For example; if parameter estimates vary from pretest to final form, items that were chosen as the best available from a pretest pool may no longer be the best choice when administered in the final form of a test.

To summarize, some degree of instability was observed for all item parameter estimates; item difficulty estimates appeared to be the most stable and estimates of the psuedo-guessing parameter the least. The item parameter estimates obtained for the aptitude test data (SAT-V and SAT-M) exhibited a higher degree of stability than those estimated for the achievement tests. Lack of stability in the parameter estimates appeared to be related more directly to differences in group ability than to time lapse between administrations.

The results of the study indicate that some degree of caution should be exercised when using parameter estimates obtained at an earlier point in time. It would seem prudent to periodically re-calibrate the items to ascertain if the parameter estimates have remained valid for a particular application and examinee population. Because lack of stability in item parameter estimates may affect applications differentially, it is suggested that prior to implementation, the affect of parameter stability on a particular application be studied and that after implementation, periodic monitoring of the item parameter estimates be carried out on a routine basis.

References

Bejar, I. I. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 1980, 17, 283-296.

Cook, L. L., Dunbar, S. B., and Eignor, D. R. IRT equating: A flexible alternative to conventional methods for solving practical testing problems. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Divgi, D. R. Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of NCME, Los Angeles, 1981 (a).

Divgi, D. R. Potential pitfalls in applications of item response theory. Paper presented at the annual meeting of NCME, Los Angeles, 1981 (b).

Green, M. S., and Divgi, D. R. The invariance of parameter estimates in three latent trait models. Paper presented at the annual meeting of NCME, Los Angeles, 1981.

Gustafsson, J. E. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement, 1979, 16, 153-158.

Gustafsson, J. E. Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 1980, 33, 205-233.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., and Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Kingston, N. M., and Dorans, N. J. The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory. Unpublished manuscript. Princeton, N.J.: Educational Testing Service, 1981 (a).

Kingston, N. M., and Dorans, N. J. The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. GRE Board Professional Report 79-12. Princeton, N.J.: Educational Testing Service, 1981 (b).

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

Lord, F. M., and Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Lord, F. M., and Stocking, M. L. Developing a common metric in item response theory. Unpublished manuscript, Princeton, N.J.: Educational Testing Service, 1982.

Petersen, N. S., Cook, L. L. and Stocking, M. L. IRT versus conventional equating methods: A comparative study of scale stability. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Rentz, R. R. Monitoring the quality of an item-pool calibrated by the Rasch model. Paper presented at the annual meeting of NCME, Toronto, 1978.

Rentz, R. R. The effects of instruction on the stability of item parameters from two latent trait models. Paper presented at the annual meeting of NCME, New York, 1982.

Ridenour, S., and Rentz, R. R. Maintaining an item bank's underlying Rasch ability scale. Paper presented at annual meeting of AERA, Boston, 1980.

Slinde, J. A., and Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Slinde, J. A., and Linn, R. L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 1979, 16, 159-165.

Whitely, S. E., and Dawis, R. V. The influence of test context on item difficulty. Educational and Psychological Measurement, 1976, 36, 329-337.

Wood, R. L., and Lord, F. M. A users guide to LOGIST (RM-76-4). Princeton, N.J.: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., and Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters. (RM-76-6). Princeton, N.J.: Educational Testing Service, 1976.

Yen, W. M. The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 1980, 17, 297-311.

APPENDIX

Table 1

SAT Verbal and Mathematical Forms and Equating Sections
Chosen for Temporal Stability Study

SAT Verbal

Form	n (items)	Administration	Admin. Date	Time Lapse (months)	N (examinees)	Mean	Standard Deviation
Y3	85	1	6/76	19	2578	34.48	16.34
		2	1/78		2549	31.37	15.86
fk	40	1	4/76	37	2879	15.08	8.19
		2	5/79		2665	15.04	8.01
fw	40	1	1/78	16	2549	14.36	8.17
		2	5/79		2700	16.38	8.06

SAT-Mathematical

Form	n (items)	Administration	Admin. Date	Time Lapse (months)	N (examinees)	Mean	Standard Deviation
Y3	59[1]	1	6/76	19	2553	24.05	13.30
		2	1/78		2455	21.48	13.74
fn	24[2]	1	4/75	14	2527	9.73	5.73
		2	6/76		2553	9.57	5.85
fx	25	1	1/78	16	2455	8.47	6.33
		2	5/79		2633	10.14	6.10

[1]Scores on SAT-M form Y3 are based on only 59 items due to a printing error in one item.

[2]Scores on the mathematical anchor test fn are based on only 24 items due to a printing error in one item.

Table 2

Achievement Test Forms Chosen for Temporal Stability Study

Biology

Form	n (items)	Administration	Admin. Date	Time Lapse (months)	N (examinees)	Mean	Standard Deviation
VAC1	100	1	1/73	52	2101	43.70	17.94
		2	5/78		3253	48.38	18.77
TAC2	100	1	1/78	16	2511	43.75	18.70
		2	5/79		3032	47.59	19.88

American History and Social Studies

Form	n (items)	Administration	Admin. Date	Time Lapse (months)	N (examinees)	Mean	Standard Deviation
YAC2	100	1	12/76	25	2120	38.73	15.13
		2	1/79		2317	37.18	15.18
AAC	100	1	12/78	18	2102	40.30	16.60
		2	6/80		2031	46.93	17.92

Table 3

Correlations and Summary Statistics for Item Parameters
SAT Verbal Forms and Equating Sections

SAT Verbal Y3
June 1976

SAT Verbal Y3 January 1978		a	b	c	Mean	S.D.
	a	.806			.857	.313
	b		.977		.265	1.426
	c			.681	.148	.054
	Mean	.883	.258	.157	n=	85
	S.D.	.298	1.320	.051	MAD=	.0217

SAT Verbal fk
April 1976

SAT Verbal fk May 1979		a	b	c	Mean	S.D.
	a	.917			.878	.285
	b		.993		.348	1.224
	c			.779	.146	.031
	Mean	.837	.364	.145	n=	40
	S.D.	.249	1.225	.037	MAD=	.0212

SAT Verbal fw
January 1978

SAT Verbal fw May 1979		a	b	c	Mean	S.D.
	a	.910			.870	.320
	b		.987		.391	1.234
	c			.420	.141	.054
	Mean	.836	.418	.140	n=	40
	S.D.	.309	1.170	.046	MAD=	.0256

Table 4

Correlations and Summary Statistics for Item Parameters
SAT Mathematical Forms and Equating Sections

SAT Math Y3
June 1976

SAT Math Y3 January 1978		a	b	c	Mean	S.D.
	a	.920			.962	.338
	b		.991		.129	1.207
	c			.844	.132	.066
	Mean	.972	.134	.133	n=	59
	S.D.	.334	1.185	.068	MAD=	.0199

SAT Math fn
April 1975

SAT Math fn June 1976		a	b	c	Mean	S.D.
	a	.771			.845	.211
	b		.993		.075	1.378
	c			.832	.114	.066
	Mean	.895	.158	.118	n=	24
	S.D.	.274	1.334	.055	MAD=	.0234

SAT Math fx
January 1978

SAT Math fx May 1979		a	b	c	Mean	S.D.
	a	.893			1.018	.263
	b		.991		.419	1.088
	c			.823	.101	.049
	Mean	1.042	.420	.110	n=	25
	S.D.	.296	1.065	.056	MAD=	.0240

Table 5

Correlations and Summary Statistics for Item Parameters
Achievement Test Forms

Biology VAC1
January 1973

Biology VAC1
May 1978

	a	b	c	Mean	S.D.
a	.814			.634	.210
b		.957		.030	1.225
c			.473	.157	.054
Mean	.668	.043	.166	n=	100
S.D.	.228	1.242	.067	$\overline{MAD}$=	.0335

Biology TAC2
January 1978

Biology TAC2
May 1979

	a	b	c	Mean	S.D.
a	.856			.677	.228
b		.967		.309	1.195
c			.472	.179	.055
Mean	.701	.285	.172	n=	100
S.D.	.247	1.201	.070	$\overline{MAD}$=	.0266

Table 5 (continued)

Correlations and Summary Statistics for Item Parameters
Achievement Test Forms

American History YAC2
December 1976

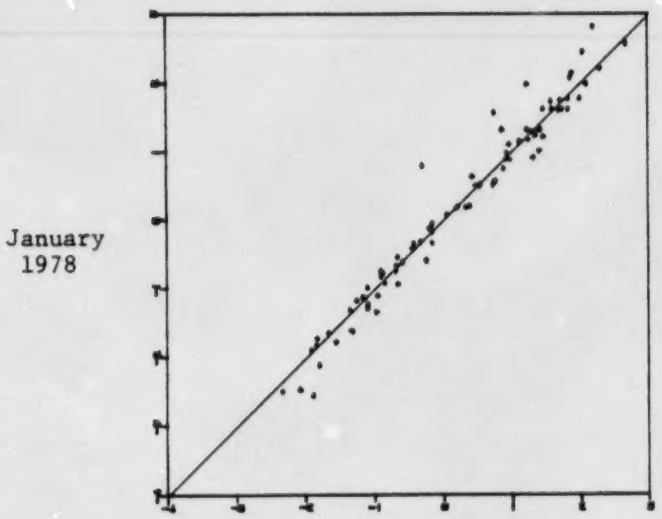
American History YAC2
January 1979

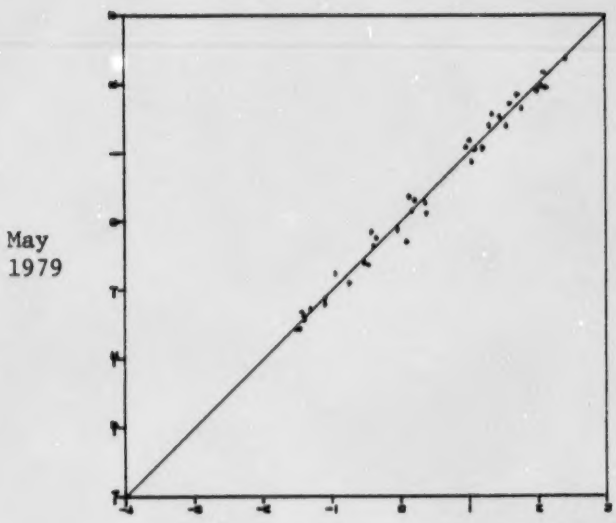
	a	b	c	Mean	S.D.
a	.746			.659	.231
b		.977		.321	1.821
c			.561	.161	.061
Mean	.629	.364	.161	n=	100
S.D.	.213	2.035	.078	MAD=	.0228

American History AAC
December 1978

American History AAC
June 1980

	a	b	c	Mean	S.D.
a	.667			.622	.211
b		.687		.211	2.553
c			.329	.173	.081
Mean	.613	.364	.161	n=	100
S.D.	.226	1.372	.065	MAD=	.0434





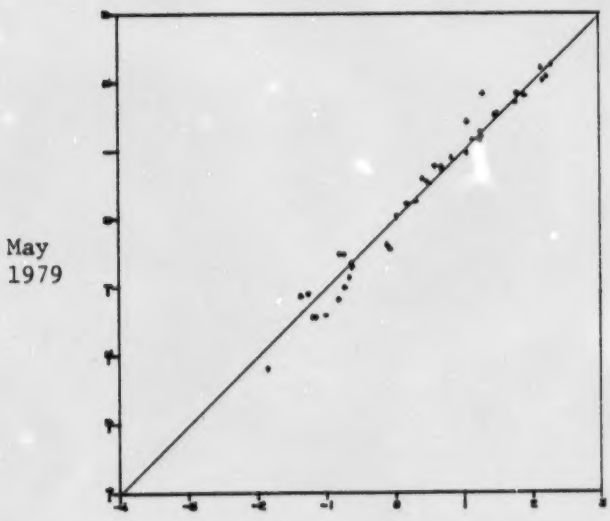


Figure 1 : Plots of item difficulty parameters (b_g) for SAT verbal forms and equating sections.

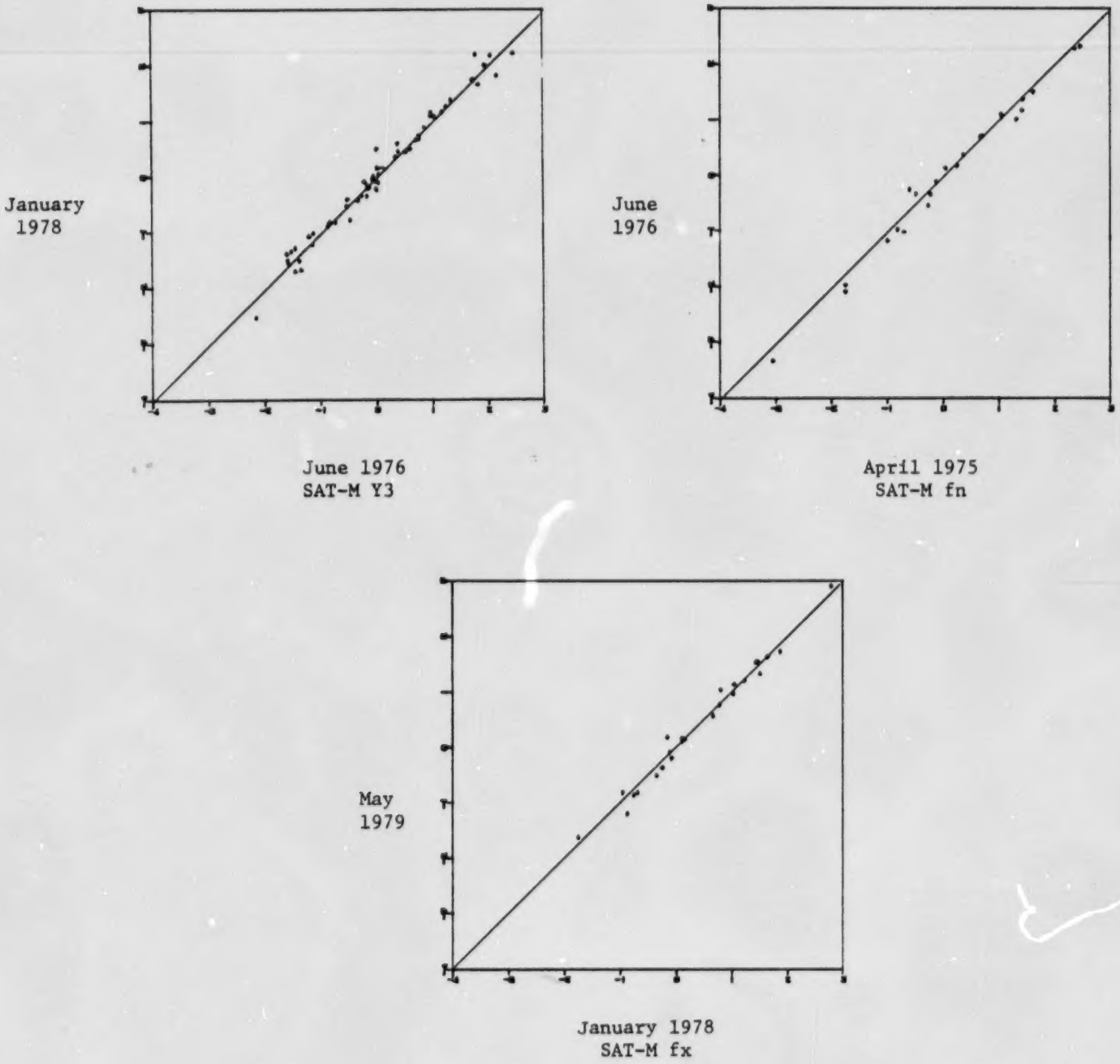
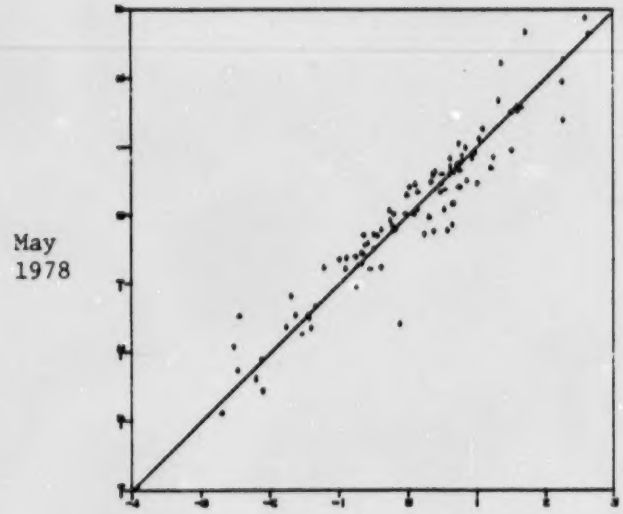
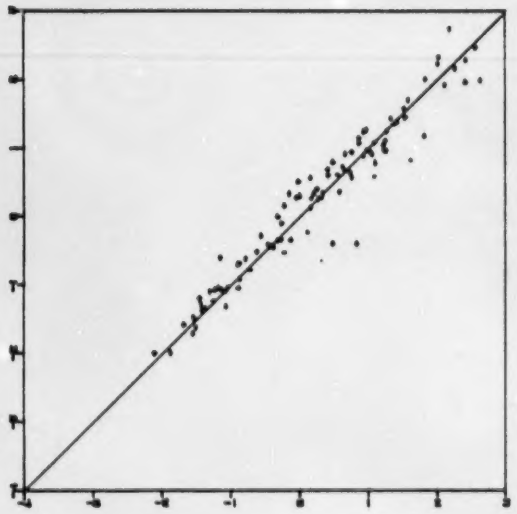


Figure 2 : Plots of item difficulty parameters (b_g) for SAT mathematical forms and equating sections.



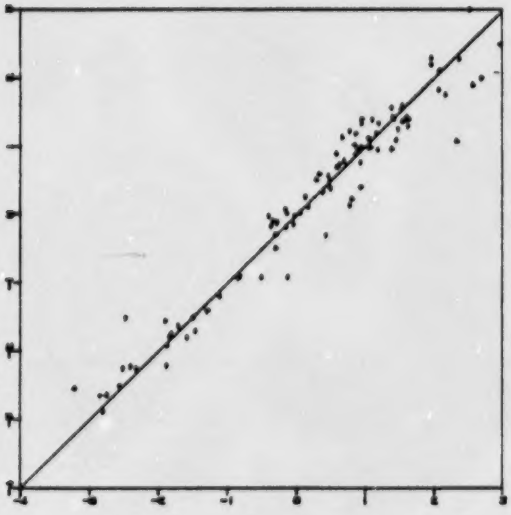
May
1978

January 1973
Biology VAC1



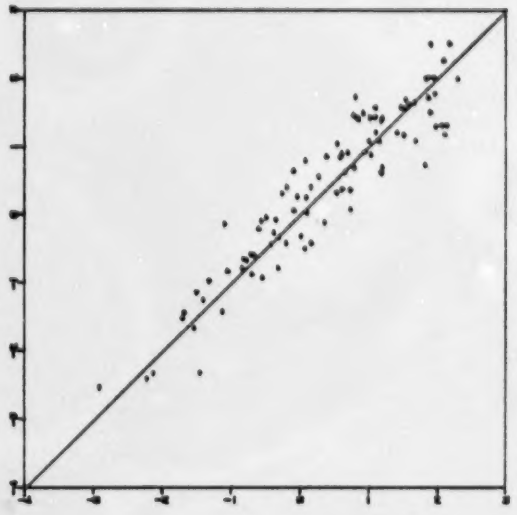
May
1979

January 1978
Biology TAC2



January
1979

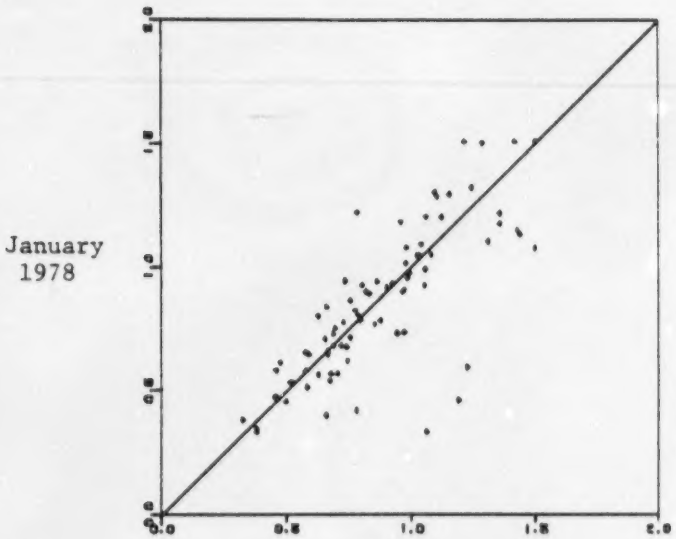
December 1976
Amer. Hist. YAC2



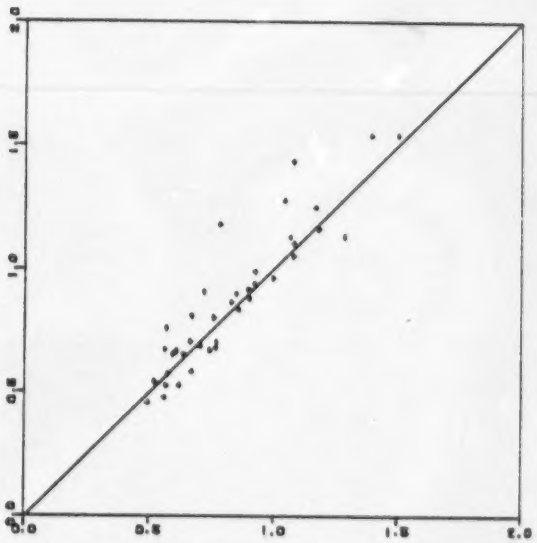
June
1980

December 1978
Amer. Hist. AAC

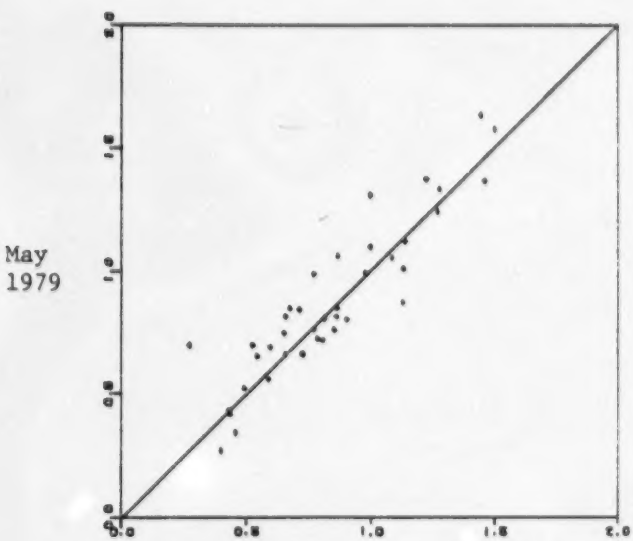
Figure 3 : Plots of item difficulty parameters (b_g) for achievement test forms.



June 1976
SAT-V Y3



April 1976
SAT-V fk



January 1978
SAT-V fw

Figure 4 : Plots of item discrimination parameters (a_g) for SAT verbal forms and equating sections.

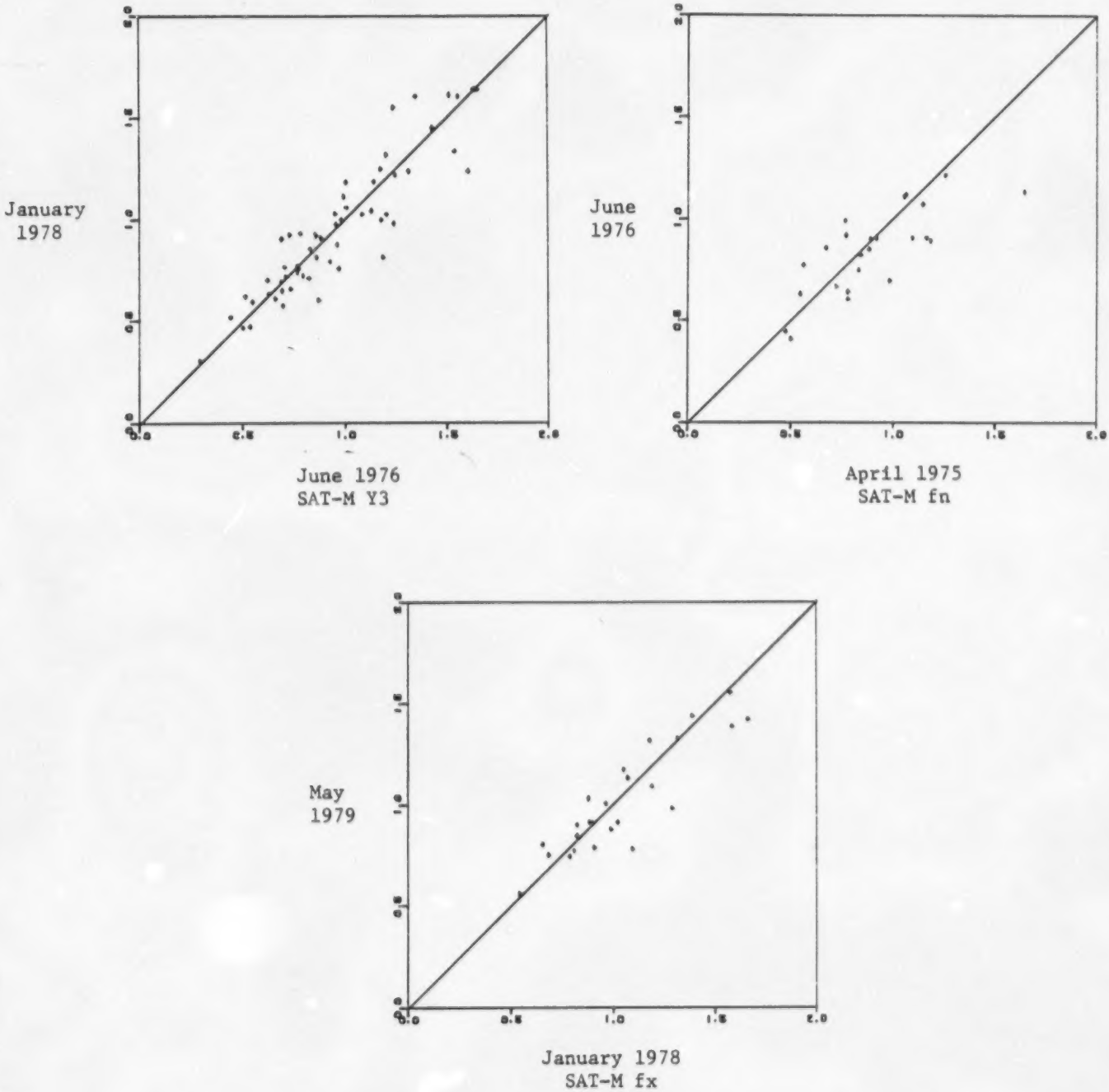
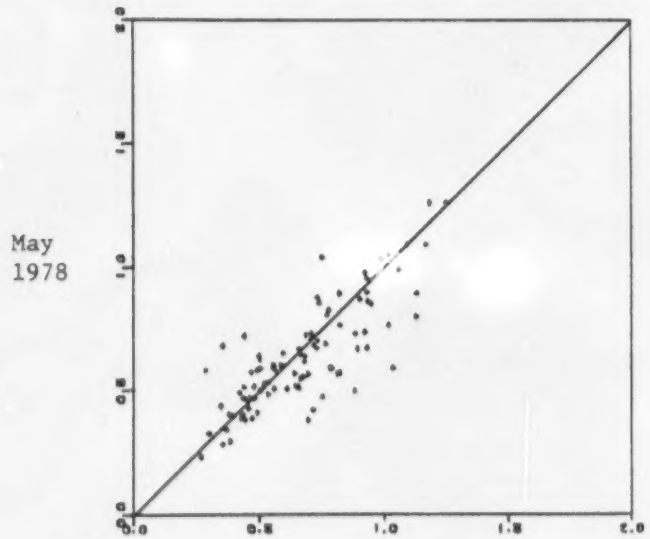
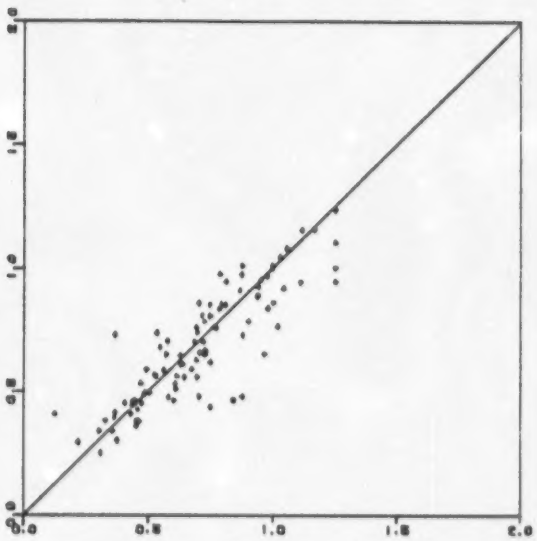


Figure 5 : Plots of item discrimination parameters (a_g) for SAT mathematical forms and equating sections.



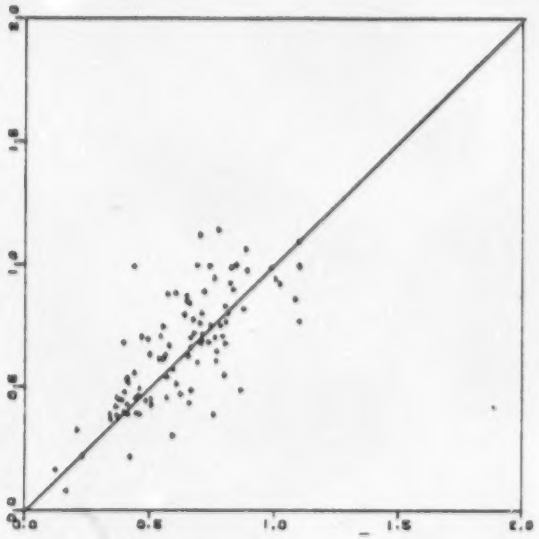
May
1978

January 1973
Biology VAC1



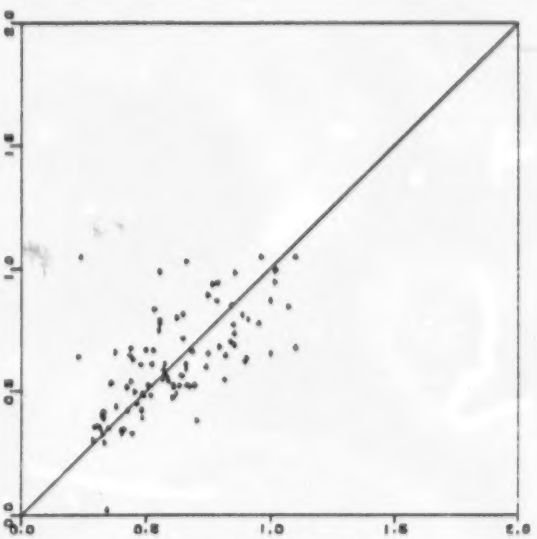
May
1979

January 1978
Biology TAC2



January
1979

December 1976
Amer. Hist. YAC2

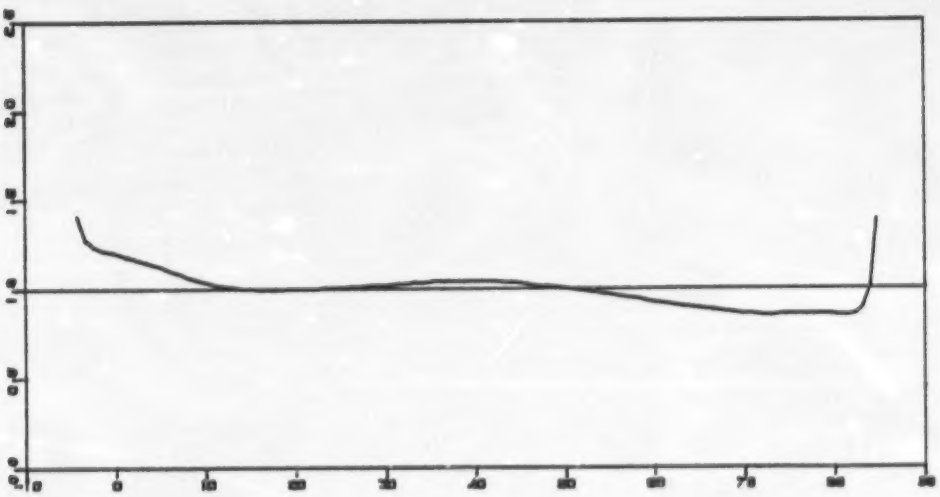


June
1980

December 1978
Amer. Hist. AAC

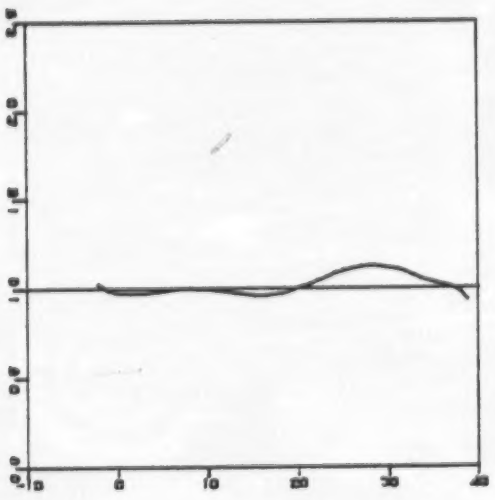
Figure 6 : Plots of item discrimination parameters (a_g) for achievement test forms.

January
1978



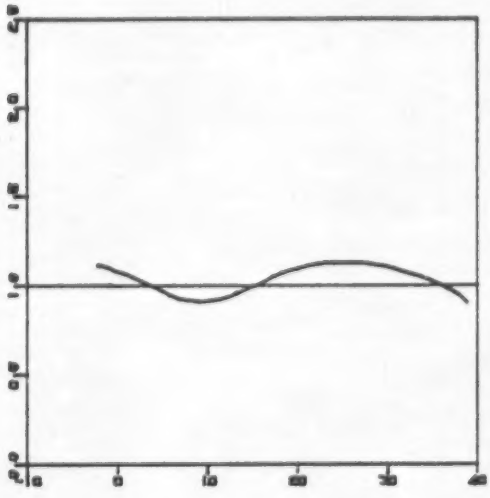
June 1976
SAT-V Y3

May
1979



April 1976
SAT-V fk

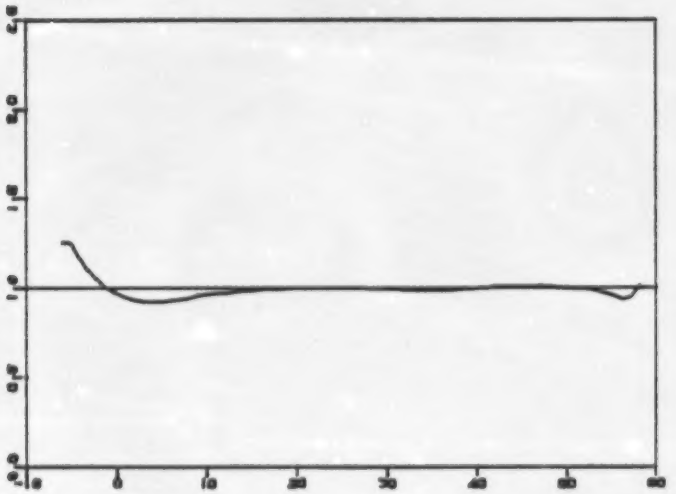
May
1979



January 1978
SAT-V fw

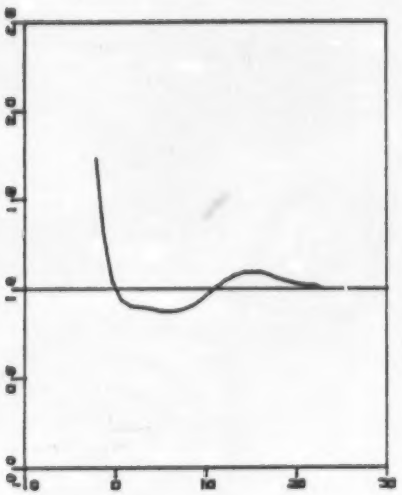
Figure 7: Relative efficiency curves for SAT verbal forms and equating sections.

January
1978



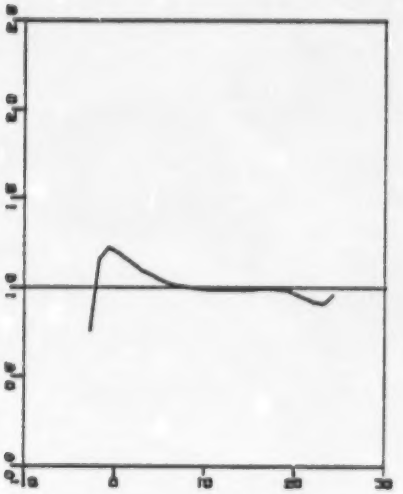
June 1976
SAT-M Y3

June
1976



April 1975
SAT-M fn

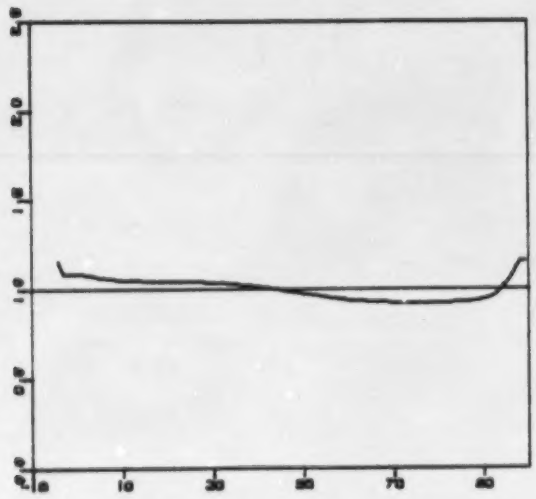
May
1979



January 1978
SAT-M fx

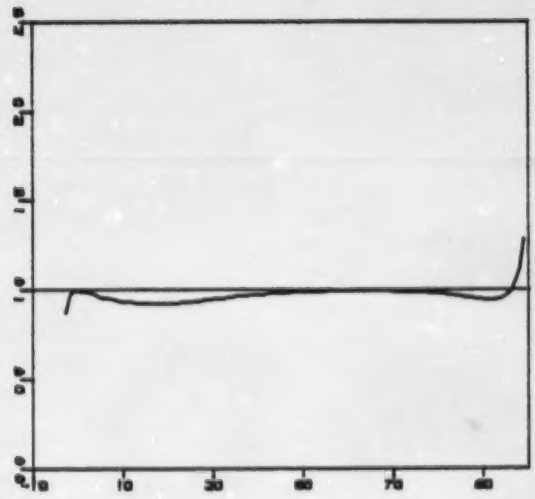
Figure 8: Relative efficiency curves for SAT mathematical forms and equating sections.

May
1978



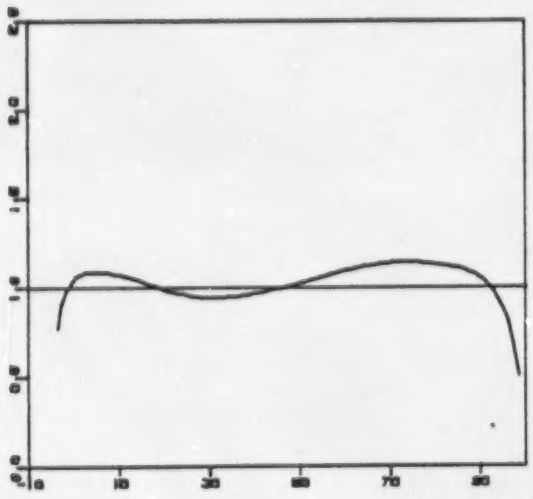
January 1973
Biology VAC1

May
1979



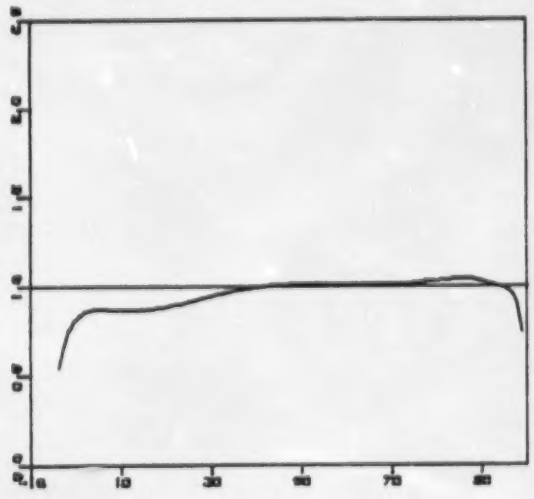
January 1978
Biology TAC2

January
1979



December 1976
Amer. Hist. YAC2

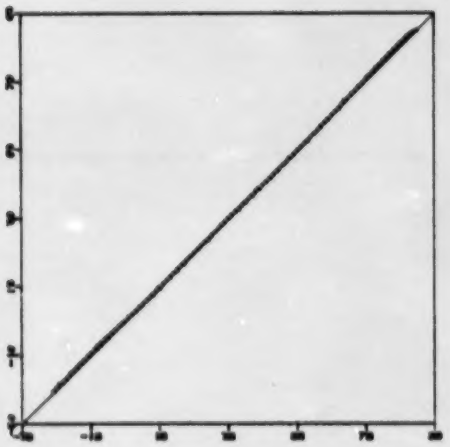
June
1980



December 1978
Amer. Hist. AAC

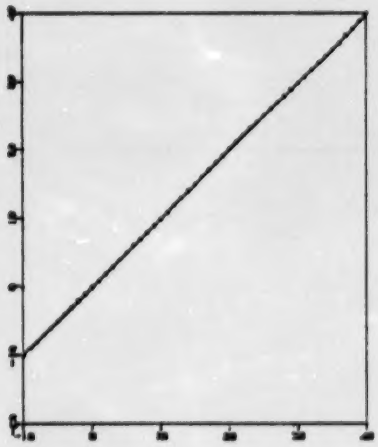
Figure 9 : Relative efficiency curves for achievement test forms.

June
1976



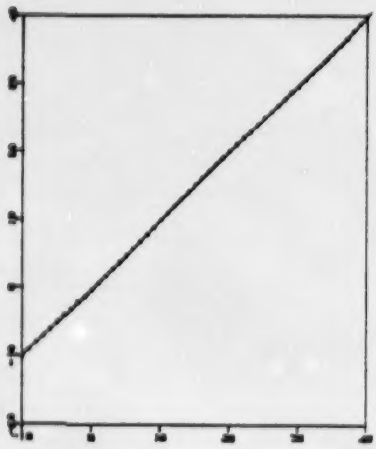
January 1978
SAT-V Y3

April
1976



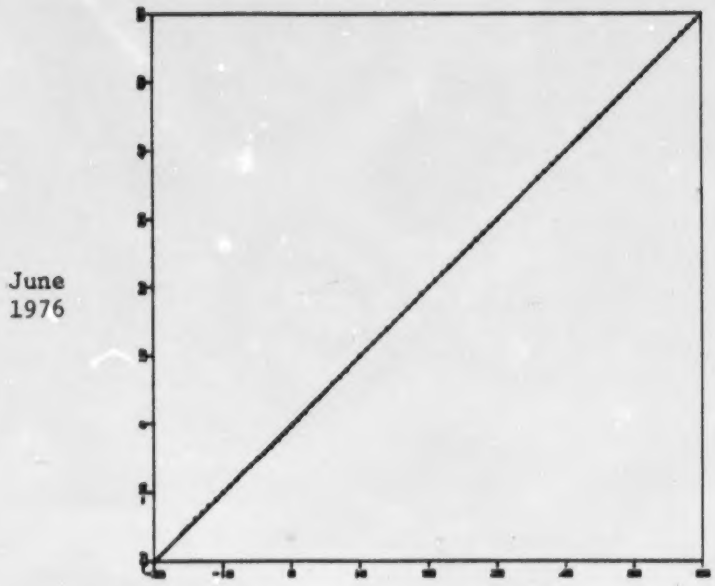
May 1979
SAT-V fk

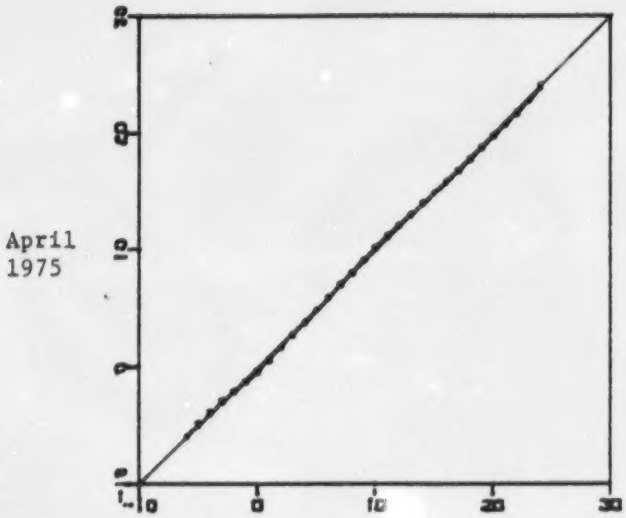
January
1978



May 1979
SAT-V fw

Figure 10 : Equating plots for SAT verbal forms and equating sections.





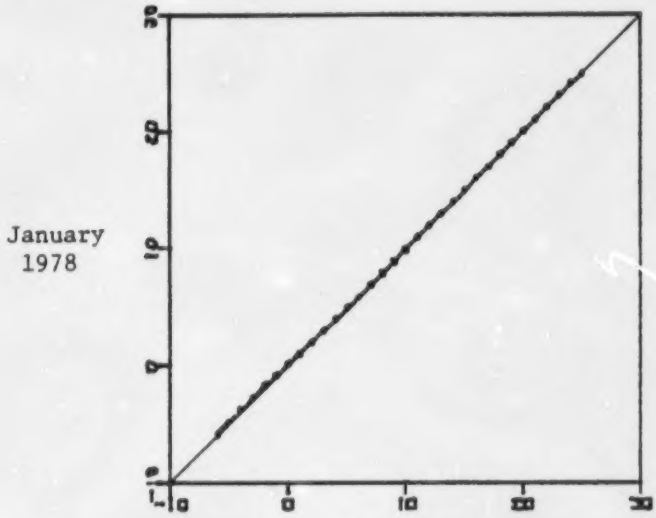
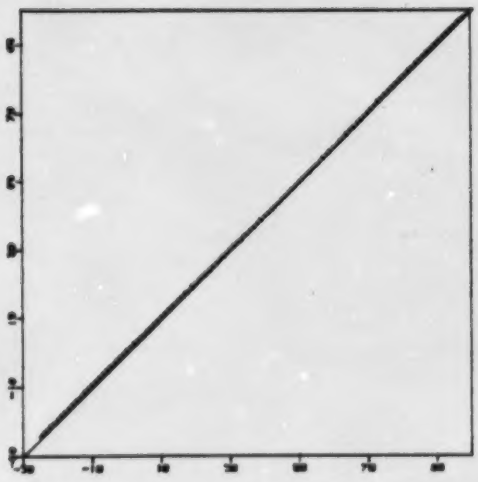


Figure 11: Equating plots for SAT mathematical forms and equating sections.

January
1973

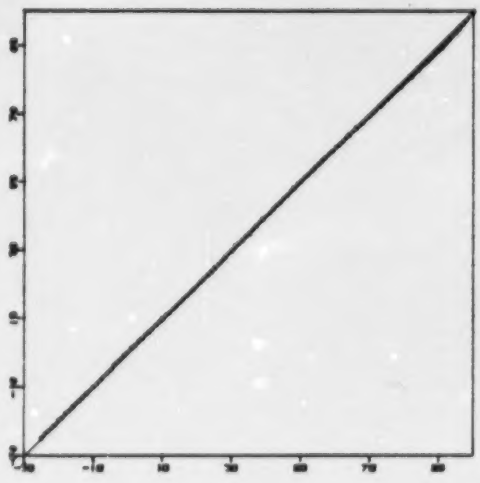


May 1978
Biology VAC1

January
1978

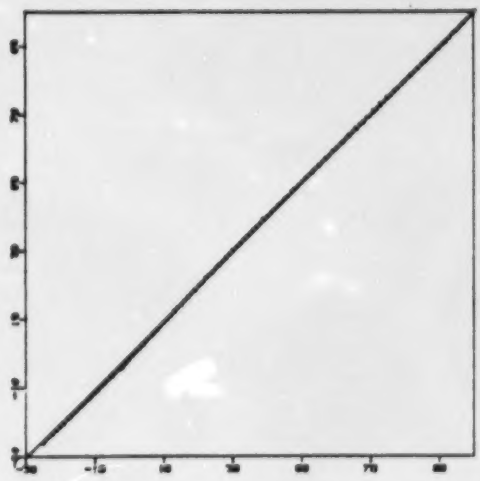
May 1979
Biology TAC2

December
1976



January 1979
Amer. Hist. YAC2

December
1978



June 1980
Amer. Hist. AAC

Figure 12: Equating plots for achievement test forms.

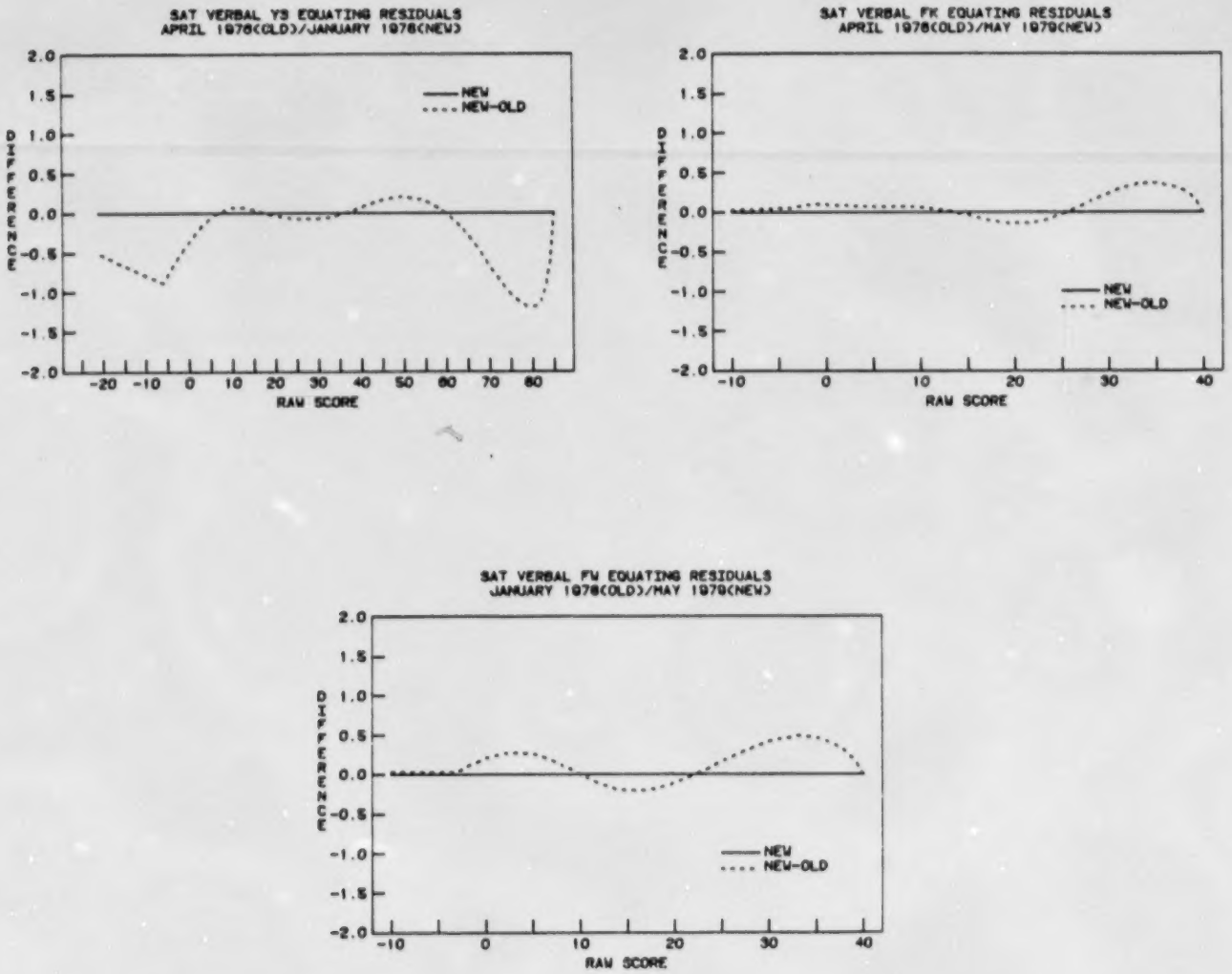
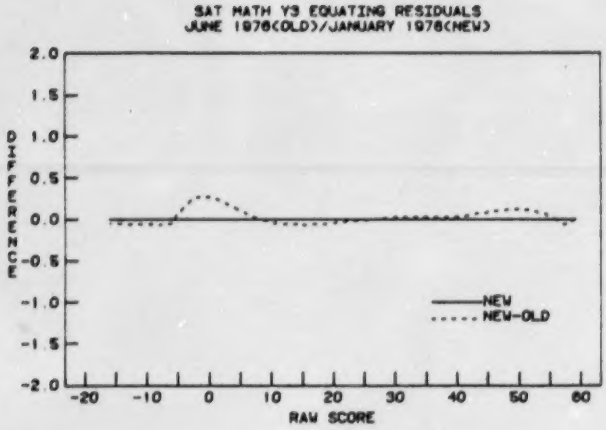
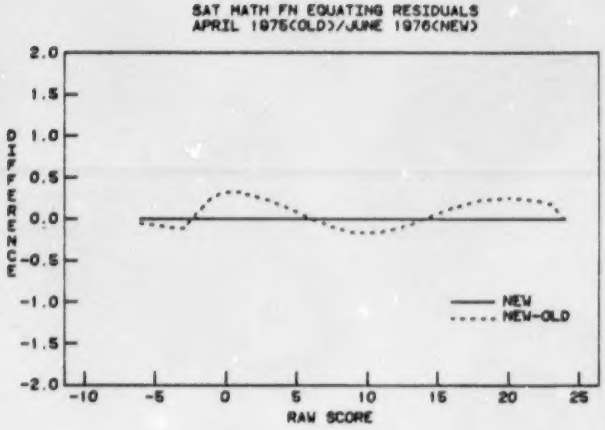


Figure 13: Plots of equating residuals for SAT verbal forms and equating sections.





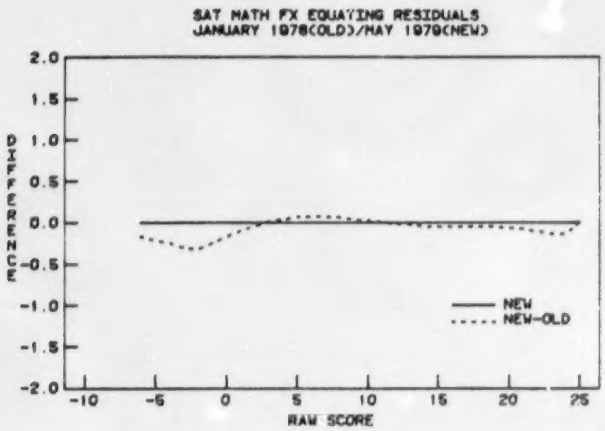
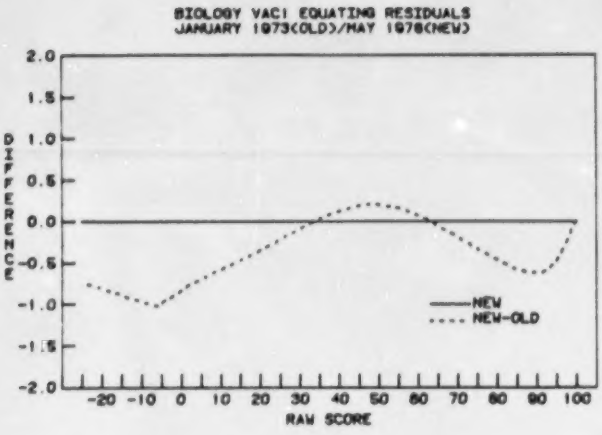
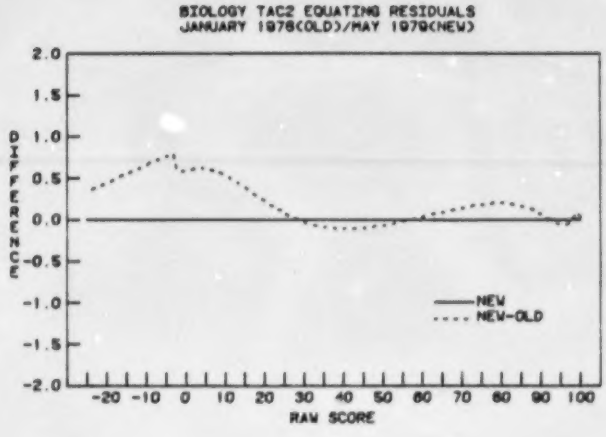
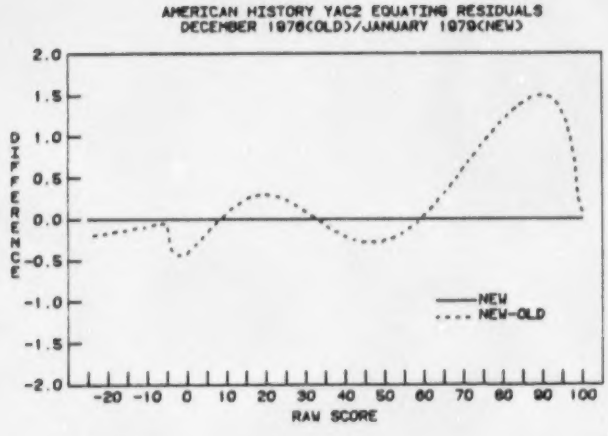


Figure 14 : Plots of equating residuals for SAT mathematical forms and equating sections.







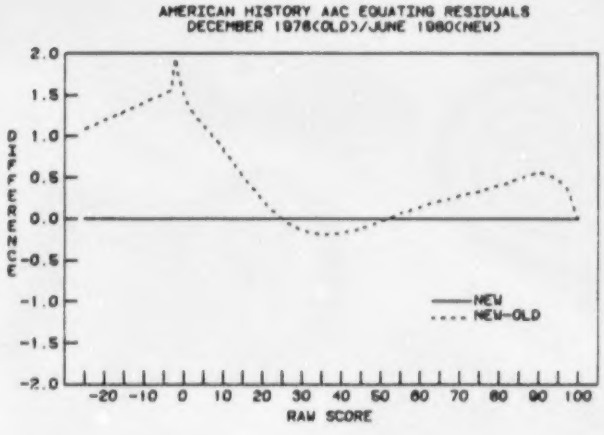


Figure 15 : Plots of equating residuals for achievement tests.